Web of Science Research Funding Information: Methodology for its use in Analysis and Evaluation

Mursheda Begum, Grant Lewison

Kings College London, Research Oncology, London, UK

ABSTRACT

This paper explains the problems associated with using the funding information on the Web of Science (WoS) to determine how research has been funded. Funding sources have been added to the WoS as searchable fields since late 2008. However a funding source can have several different names and can appear with different conventions for abbreviation, punctuation and format. We needed to have one name for each source and it was convenient for names to have the same format. We therefore use a coding system, which is described in detail. It identifies most individual funders, and gives their country and sector (public, private-non-profit, commercial or international). Some of the commercial funders listed are false positives as they have paid the author(s) for unrelated work. In addition, acknowledgements implicit from the addresses also needed to be coded for public, charity and commercial laboratories. These two sets of codes were then combined to permit a detailed analysis of funding for each country by means of two Excel programs.

Keywords: Bibliometrics, Explicit funding, Implicit funding, Conflict of interest, Acknowledgements.

INTRODUCTION

Information on the funding of research, especially biomedical research, is valuable for several reasons. First, the funders of research have an interest in learning about the papers published with their support so that they can see what has been achieved with their money. There is, of course, a huge literature on how such outputs can be evaluated, but clearly the first task in any evaluation is to list the outputs. Thus in the 1970s, several of the US National Institutes of Health commissioned CHI Research Inc. to search the biomedical literature for papers that acknowledged their support.[1] Much later, in the 1990s, the Wellcome Trust set up the Research Outputs Database (ROD) as a "club" project to examine all UK biomedical papers in the Science Citation Index and Social Sciences Citation Index© Thomson Reuters, some 30,000 per year, in London libraries and record their acknowledgements. This enabled lists of papers to be issued to the Trust and to the club members.[2,3]

***Address for correspondence:**

Grant Lewison , Kings College London, Research Oncology, Guy's Hospital, London, SE 1 6RT, UK

Email: grant.lewison@kcl.ac.uk

Phone no: +44-(0)20-7848 6742

Access this article online

Official Publication of Phcog.Net Bringing Medicinal Plants Researchers Together

Website: www.jscires.org

DOI: 10.5530/jscires.6.2.12

The second reason for the interest in funding data is that it has now been established that papers with more acknowledged funding sources tend to be more highly cited,[4–6] and that this effect swamps the supposed benefits of multiple addresses which are actually *negative*, at least up to about five addresses.[7,8] An evaluation of groups of papers that does not take account of funding information is thus ignoring a major factor influencing the likely citation scores.

A third reason for the study of financial acknowledgements is that the results may be swayed by the source of funding. This is a separate concern to that expressed in our earlier paper on conflicts of interest,[9] where individual authors report receipt of money from companies for unrelated work. Research that acknowledges commercial funding has a clear bias to positive results,[10] so it is important for readers to be able to discount the claims made if a trial has been sponsored by a company. Not all such sponsorships should necessarily be regarded with suspicion. Many papers state that the sponsor had no role in study design or analysis, or the decision to publish the results.

Finally, acknowledgements of support may point researchers to sources of funding for similar work that they were hoping to carry out. In many countries of western Europe and North America, there are so many research funding organisations (as we have discovered) that it is unlikely that researchers would be aware of all of them.

Since late 2008 the Web of Science (WoS) has routinely collected data on financial acknowledgements where they are recorded in papers. This has only been for the Science Citation Index Extended until 2015 when acknowledgements in social science journals began to be covered. Moreover, coverage is only where the text of the acknowledgement is in English; this includes the large majority of articles and reviews, and some Chinese ones, but excludes papers in other European languages.[11] There are three searchable fields:

- the full acknowledgement text, FT;
- the list of funding agencies, FO; and
- the grant number, FG.

The information can be downloaded to file with the other bibliographic data. When this is done, the acknowledgement text is in a column that is now headed FX and the funding agencies are listed in a column that is now headed FU. The names of the funding agencies are separated by a semi-colon and a space and the grant number is attached to the name of the funding agency in square brackets. (Sometimes the name of the funder is not given, but it can often be deduced from the grant number, if it has distinguishing letters, for example the codes used by the different institutes of the US National Institutes of Health). The different funders can therefore easily be separated out, and the grant numbers removed, if desired.

However, the names of the funding agencies are those given by the paper authors, and are not standardised in any way. Thus, we have found *several thousand* different formats for the European Commission, which not only has a number of programmes but each of them is named in a large variety of ways. It thus became necessary when the ROD was set up to devise a coding system in which to record the funding acknowledgements on the papers. This was originally a four-part code:

A trigraph (three-character code) connoting the individual funder, *e.g.,* MRC = UK Medical Research Council;

A digraph (two-character code) connoting the sector and sub-sector of the funding agency. These are listed in Table 1;

Another digraph connoting the country of the funding agency. EU was used for European ones, and XN for international ones (such as the World Health Organization, WHO).

Table 1: Table of sectors and sub-sectors for research funding agencies with codes.

Government GOV		**Private-non-profit PNP**		**Industry INDY**	
Code	Sub-sector	Code	Sub-sector	Code	Sub-sector
GA	Government agency	CH	Collecting charity	BT	Biotech company
GD	Government department	FO	Endowed foundation	IN	Industrial (non-pharma)
LA	Local authority	HT	Hospital trustees	IP	Pharma company
		MI	Mixed (academic)		
		NP	Other non-profit		

A single character connoting the type of support, such as a project grant, a personal fellowship, provision of equipment, provision of consumables such as drugs, or travel grants. This character has not been used since 2003 when the ROD project ended.

The trigraph codes were listed in a thesaurus that was used by our recorders to note the identity of the many thousands of different funders that were acknowledged on the UK biomedical papers. This is the "old" thesaurus; as the names of acknowledgees are encountered and given tripartite codes, they are successively added to a "new" thesaurus which contains many variants of their names.

Some types of funders were not specifically acknowledged but were implicit from the papers' addresses. These were government laboratories (including regional or local authority ones in some countries), charity units and commercial companies, but not endowed foundations. Funding bodies of these three sectors featuring among the addresses were regarded as "implicit" funders, as opposed to the "explicit" funders listed in the acknowledgement section of the paper (and recorded in the WoS as FO). However, many authors ignore these implicit acknowledgements even though they certainly indicate the sources of funding for the research, if they are not already among the explicit acknowledgements. Whilst funders request acknowledgement of their contribution we are not aware of any cases of review and / or enforcement of this with grantees.

As discussed previously,[9] some of the funders listed in the FU column in the downloaded file were false positives because the companies had paid one or more of the authors for unrelated work. This means that for some of the papers the FU entry has to be redacted. The earlier paper described in detail how the relevant papers are identified from the acknowledgement text FX and how the FU entry is changed.

The objective of this paper is to describe a procedure whereby the names of research funding agencies can be unified by means of a coding system so that their contributions to a research field can be analysed. We show that there are huge numbers of different funders and that their names are written by authors in a multitude of different formats. Our system allows for them to be appropriately classified, based on several decades' experience in this process. We then describe how these codes can be combined and the analysis of funding completed by means of two special programs so as to show the funding sources for any given set of papers, such as those from a single country.

METHODOLOGY: EXPLICIT ACKNOWLEDGEMENTS

For a given set of research papers, it is relatively easy to collect the individual funders acknowledged in the FU column after redaction and assemble them into a single column of an Excel spreadsheet. The grant number in square brackets is removed, but some acknowledgements only give the grant number, and not the name of the funding source, and this is retained as it can often identify the source of funding. Then a program allows the numbers of each acknowledgee to be determined, with them being listed alphabetically. Each one is now looked up in our "new" thesaurus of funding organisations which, unlike the "old" thesaurus used by the ROD recorders, includes all the funder name variants so far encountered, with their codes: some may also appear in the "old" thesaurus. These acknowledgees are then marked with a sign in column B of the spreadsheet to show that they exist in the new thesaurus. This thesaurus is proprietary, and currently contains over 100,000 entries. It is constantly being updated to take account both of new funding bodies, and changes in status of ones already there.

However, inevitably the large majority of acknowledgees will not yet be coded. It is then necessary to work through the list, a long and somewhat tedious job, to assign codes to almost all of them (a few appear to be acknowledgements to named individuals, which have been wrongly recorded as sources of funding, or arise from errors in punctuation, or are too vague to be clearly identified: these are given "NO CODE"). Many are obvious variants of funding organisations already coded, and these codes can be copied and pasted down. These acknowledgees, without the mark in column B, will subsequently need to be added to the new thesaurus with their assigned codes. (As the coding of funders is hardly an exact science, it sometimes happens that we learn more about an organisation and need to change its code in both the "old" and "new" thesauruses.) But there will also be a large number of new funding organisations that need to be coded.

The system of codes devised for the ROD in 1993 was sufficient for 26 x 26 x 26 = 17,576 codes AAA to ZZZ. We supposed that that number would suffice to record all funders of biomedical research. However it soon became apparent that we had grossly under-estimated the population of funders, not least because many of the UK papers (with which the ROD was concerned) were co-authored internationally and so we needed to accommodate foreign funders, especially ones in North America and continental Europe. We started to assign codes to the ones we encountered, but it was clear that it was a losing battle to give unique codes to every funder that we found, so we decided to assign "generic" codes to funding organisations in the main UK partner countries. These consisted of the letter X, Y or Z, followed by a single digit to show the country: X0 = Netherlands, X1 = USA, X2 = Germany, and so on; and another single digit to show the sector and sub-sector, as shown in Table 2. This is not strictly needed as the three-part code includes the sub-sector and country for all funders, but it means that the generic code has three characters, so that all codes have the same length.

Funding organisations from countries not on the list of 25 selected UK partner countries were given a code indicative of their continent: Z0 = Europe; Z7 = Africa; Z8 = Asia and Z9 = Latin America. However, there was no loss of information because the other two parts of the code showed the sector (Table 2) and the individual country through its ISO code.

Inevitably, many of the acknowledgees needed to be researched in order to determine their country and sector. Fortunately, the large majority of them have a website, and this can be inspected to determine their location and what they do, sometimes with the help of the Google translate function. Some funding acknowledgements are to a generic organisation that could come from any country, such as "Ministry of Health". If the beneficiary of this grant is indicated (usually by their initials), then their address can be sought in the C1 column of the spreadsheet, in which

Table 2: Single-digit codes for sector and sub-sector of funding organisations assigned generic codes.

Sector	Code	Sector	Code	Sector	Code
Collecting charity	1	Industrial company	5	Other non-profit	9
Endowed foundation	2	Pharma company	6	Biotech company	B
National government	3	Local authority	7		
Hospital trustees	4	Mixed (academic)	8		

the names of authors are matched with their addresses, and hence the country of the ministry determined.

Some acknowledgements in the FU column of the spreadsheet are to the same funder, often because different partners both received grants from the funder and may have written their acknowledgements separately before they were put together for the paper. This occurs both with commercial companies and other funders, such as the European Commission. The duplication of codes that results (because the acknowledgees are treated individually) is not necessarily wrong, as there have effectively been two (or more) grants. Duplication of codes can also occur when there are several funders with the same generic code (e.g., several US biotech companies); again this is not wrong because if they had been given individual codes they would not have been duplicated.

Although many different types of organisation have the legal status of charities, we only use the code CH for those who solicit money from the public and whose main business is the support of research. For example, in the UK the Medical Research Council has this legal status, and can accept charitable donations, but its main source of funding is government, so it is classed as a government agency, coded as GA. "Charities" include some organisations called "foundations" but which are really collecting charities named for someone who died from cancer or other disease, and are set up by their relatives or friends. However, patient support groups providing nursing care or information are coded as NP, as are professional associations and organisations with quite different main interests which also provide some support for research. These include business social clubs, such as the Kiwanis, Lions and Rotary, which are international but have local chapters. We have come across some surprising sources of medical research funding, including sporting clubs such as the immensely rich Hong Kong Jockey Club, and clubs for owners of dog breeds, and a few of them are listed in Table 3.

We have found many true foundations, usually named for individuals, and these are coded FO. These are often endowed by industrial magnates and give grants from their own resources, and they are particularly numerous in Scandinavia. Thus we have encountered over 230 in Denmark alone, of which a small sample is shown in Table 4.

In Italy and Spain, many banks were originally founded as social enterprises and still retain a philanthropic remit to assist their locality; this is discharged through individual foundations which we code FO and not IN. Legacies are coded similarly unless they are associated with and dispersed by another body, such as a professional association (e.g., the Royal College of Surgeons), when they are coded

Table 3: List of some organisations that support medical research although it is far from their main focus.

Research funding organisation	Code
French Federation of Ice Sports, 41-43 rue de Reuilly, 75012 Paris	X79-NP-FR
Gilbert Ballet Association, Limoges, France	X79-NP-FR
Global Forest Science, Palm Desert, CA, USA	X19-NP-US
International Association for Plant Taxonomy, Bratislava, Slovakia	Z09-NP-XN
Iowa Pork Producers Council, Urbandale, IA 50322, USA	X19-NP-US
MAWAS in Palangkaraya (wild orangutans habitat), Palangka Raya, Indonesia	Z89-NP-ID
Nautilus Watersports, Port Vila, Vanuatu	Z85-IN-VU
North American Electric Reliability Corporation (NERC), Atlanta, GA, USA	X19-NP-US
Northlink Ferries, Stromness, Orkney, KW16 3BH, UK	UK5-IN-UK
PHANA (Ancient prehistory in the North-eastern Alentejo), Portugal	Z09-NP-PT
Red Bull GmbH, Fuschl am See, Austria	Z35-IN-AT
Southern Tree Breeding Association, Mount Gambier, SA, Australia	Z59-NP-AU
Suffolk Sheep Society, Ballymena, Northern Ireland	UK9-NP-UK
Xstrata Coal, Zug, Switzerland (now part of Glencore plc)	Y25-IN-CH

Table 4: List of some Danish endowed foundations that support medical research, or have done so. Note that many are endowed by both husband and wife, and some give the husband's job title. The names are sometimes in English, sometimes in Danish, and sometimes in a mixture.

Danish foundation name	Code
Fru Astrid Thaysens Legat for Laegevidenskabelig Grundforskning	ATY-FO-DK
Carlsberg Foundation (Carlsbergfondet)	CBF-FO-DK
Dagmar Marshalls Fond	DMF-FO-DK
Foundation of Aase and Ejner Danielsen	EJD-FO-DK
Fabrikant Einar Willumsens Mindelegat	EWM-FO-DK
Gerda and Aage Haensch Foundation	GAH-FO-DK
Grosserer L. F. Foghts Foundation	GSF-FO-DK
Director Ib Henriksens Foundation	IBH-FO-DK
Krista and Viggo Petersens Foundation	KOV-FO-DK
Aage and Johanne Louis-Hansen Fund	LHN-FO-DK
H. Lundbeck foundation	LUK-FO-DK
A.V. Lykfeldts fund	LYK-FO-DK
A. P. Møller and Wife Chastine Mc-Kinney Møller Foundation	MMO-FO-DK
Foundations of Novo Nordisk	NID-FO-DK
Esper and Olga Boel Foundation	OEB-FO-DK
foundation of ENT doctor Hans Skouby and wife Emma Skouby	OSB-FO-DK

Grosserer Vald. Foersom og hustru Thyra Foersoms Fond	VTH-FO-DK
Else & Mogens Wedell-Wedellsborgs Foundation	WWD-FO-DK

to that body because it, rather than the testator, makes the decision on the allocation of funds. There are also foundations set up by other large companies, including many of the big pharma companies which are big enough to merit their own trigraph codes.

There are many acknowledgements to research institutes, which are funded by a variety of sources, and it is presumed that they have some general funds available to support research, often internally, but sometimes extramurally. Our practice is to code to their stated main funders if there are four or fewer, but if more than four, the institute is just given a single NP code. Some acknowledgements are to one funder "through" a second one, perhaps a university or research institute. We usually code the main source of funds, unless it appears that both parties are distinct and contributed to funding support (e.g. US National Cancer Institute designated centres are based in different institutions, usually universities, and such centres are coded as both the NIH (as it provides substantial financial support) and also the host institute (as they may also have other funding sources).

We code industrial companies (including state-owned industries) as BT, IN or IP. In principle, IP is given to companies that have a licence to manufacture and sell drugs, and BT to young firms that would like to do this, and have some new drugs under development. IN is used for non-pharma companies, including many small commercial firms providing specialist services through their laboratories, such as pathology, or are makers of medical devices. University spin-out companies are coded BT or IN as appropriate. All these codes will often be duplicated from the codes given to organisations implicitly funding the research as shown by their addresses, and these duplicates are subsequently removed (see below).

Much research is funded by government, either national, regional or local. We use three sectoral codes: GD for departments that are controlled by ministers, GA for agencies that are at least nominally independent of ministerial control, and LA for regional, county or city funding agencies. In some countries, such as Belgium, Italy and Spain, these are playing an increasing part in the funding of research. Although it appears from the above that each funder can be given a proper code, either an individual or a generic one, there are several possible problems. The first is that the landscape of commercial companies is ever-changing: some buy others (particularly big pharma buys small biotech companies), and sometimes divisions of a major company are hived off. If the latter, then a new code has to be given to the hived-off funder; if the former, then our practice is to code to the parent company on the grounds that it has purchased the intellectual property of the acquisition and so should be credited with the latter's research portfolio. This is often difficult as big pharma companies may have hundreds of subsidiaries with different names and they should really be coded to their parent. We have also found acknowledgements to companies that have subsequently gone out of business; these usually only merit a generic code for the country.

Name changes and takeovers occur also in the governmental and private-non-profit sectors. Thus governments often change the names of departments and their functions, and set up (and occasionally close) agencies to fund research in selected subjects (genetics and nanotechnology have been fashionable recently). There is a problem with government agencies, GA, that have individual funders and even sub-funders. The most prominent is the US National Institutes of Health (NIH), which is sometimes acknowledged thus, and sometimes with the name of one or more individual institutes, which are the actual funding sources. The Netherlands has a science funding agency, Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), and this has eight different divisions, some of which have individual trigraph codes in our system. Some charities with similar names may be different (and perhaps rival) organisations: in Sweden there is a dementia association, an Alzheimer association, and another one that collects for both. Charities also merge and at least in the UK there have been several mergers during the last 23 years (since the ROD was established), notably the Imperial Cancer Research Fund merging with the Cancer Research Campaign (originally the British Empire Cancer Research Campaign, when the UK still had an empire) to form Cancer Research UK, the biggest collecting medical research charity in the world, with an annual turnover now in excess of £ 620 million.[12] This means that the codes in the new thesaurus may need changing at some future date, and it is therefore necessary to keep a record of changed codes.

Some acknowledgements are so vague or difficult to interpret (for example, they may be "Ministry of Health" without specifying the country, or an acronym) that they cannot readily be coded. Help is often available from the original paper, both from the acknowledgement full text and from the addresses on the paper. These may indicate the country of the acknowledgee, or the organisation whose initials form the acronym. In order to identify the relevant paper, the FU column (or FU-R if another name has been given after the redaction) has to be filtered. However, some acknowledgements are more than 255 characters long, and if the string being searched is beyond this limit, the filter function in Excel will not reveal it. We used a special program for this purpose, originally written by Judit Bar-Ilan of the Bar-Ilan University in Israel in

1999 and still in use. Some of these acknowledgements on a particular set of papers may be identifiable uniquely, but it would not be appropriate to use the same codes (e.g., for "Ministry of Science") for other bibliometric studies as their names might then refer to organisations in different countries. Such acknowledgees are marked with a symbol to indicate that the codes assigned should NOT be copied back to the thesaurus but used only for the current campaign.

From time to time, we encountered a funding organisation with many acknowledgements that had only been given a generic code, but seemed to merit an individual code. This code was selected from the gaps in the old funding thesaurus, with initials that matched where possible the name of the new funder. A note is made indicating that this was a new code and showing the previous generic code. This needed also to be changed in the new thesaurus, and in the list of funding bodies, with a search on both the generic code and the names of the acknowledgees.

Although it is obviously desirable to allocate codes to all the acknowledgees, bibliometrics is not like particle physics where theory predicts experimental results to one part in 1010 so it will be inevitable that a few codes will be incorrect. It is probably better to accept a certain error rate in order to complete the assignment of codes in a reasonable time.

METHODOLOGY: IMPLICIT ACKNOWLEDGEMENTS

Implicit funding acknowledgements are taken from the addresses, which are collected together in a single column of an Excel spreadsheet. From previous work, we have developed an "address thesaurus" that lists all the addresses found so far, and codes for about 15% of them. [At the time of writing it contains nearly 600,000 different addresses.] We have developed rules for the identification of addresses that should be ignored, i.e., not coded, and for those that need to be coded. These are based on particular strings within the first item of the address, which is the name of the institution. Many addresses have already been met during our work on other projects, been listed in our "address thesaurus" and were either given codes or given "no code". In particular, since no code is given to hospitals, research institutes or universities, the presence of one or more of the strings listed in Table 5 means that an institution is given "no code".

Table 5. List of strings in the institutional name in the address field of a biomedical paper that indicate that it should NOT be coded.

Academic	Hospitals & Clinics		Streets
ACAD-	ALLERGY	KLIN	-AVE or -AV
AKAD	AZ-	KRANKENH	-BLVD
COLL	AZIENDA	MED-CTR	-CLOSE
FAC-MED	BIOMED-	NHS-	-COURT
KCL	CHARITE	OSPED	-CRESCENT
MED-SCH	CHU-	SJUK	-DR
SCH-MED	-CLIN	SPITAL	-GARDENS
UCL-	HOP-	SYGEHUS	-LANE
-UCL	-HOP	-TRUST	-PL
UCLA	HOSP	ZIEKENHUIS	-RD
UNIV	INFIRM		-RUE-
			-ST or -STR
			-VIALE-
			-VIA-
			-WALK
			-WAY

On the other hand, some strings given after the names of commercial companies show that they should be coded; these are listed in Table 6. These strings are not always present, and we have found it necessary to search the addresses for the presence of individual company name strings in order to ensure that they have all been appropriately coded. This has involved us surveying the mergers and acquisitions landscape for pharmaceutical and biotech companies in some detail in order to keep up with changes in ownership. Some companies sponsor university units or departments, and these are credited to the company even though they are in academia. Because the WoS system involves the use of an address thesaurus, the addresses are actually much more uniform than the names of funding organisations, so many of the institutions will already be in the "address thesaurus" and either coded or given "no code".

Although in general, hospitals and universities are not coded, some addresses also contain the names of state funding bodies: in France, CNRS and INSERM; in Germany, DFG; in Italy, CNR; in Spain, CARLOS III; and in the UK, NIHR and MRC. The presence of these national funders (and some others) should over-ride the "no code" normally given to hospitals and universities. This is because these funders support research teams in universities and institutes, and therefore the funding for the research described in the paper derives from them and not from the university or institute. Some of the sources of support for cancer research are the cancer registries that have been set up in many countries, either on a regional or a national basis. The ones that record all cancer cases are to be coded, but many of the ones recording specialist cancers are non-profit organisations and are not coded.

Although use of the VLOOKUP function in Excel can code many addresses, and advanced filtering with the terms in Tables 5 and 6 can code many more, there will

Table 6: List of strings given after the names of commercial firms showing that they are limited companies, and should be coded (BT, IN or IP). Note: sometimes the strings are given with intermediate stops, e.g., A.B. and sometimes not, e.g., AB

String	Which stands for	Country
A.B.	Aktiebolag	Sweden
A/S	Aktieselskab	Denmark
AG	Aktiengesellschaft	Germany/Austria
ASA	Allmennaksjeselskap	Norway
B.V.B.A	Besloten vennootschap met beperkte aansprakelijkheid	Netherlands
GmbH	Gesellschaft mit beschränkter Haftung	Germany/Austria
Inc	Incorporated	United States
K.K.	Kabushiki Kaisha	Japan
Ltd	Limited Company	United Kingdom
N.V.	Naamloze vennootschap	Netherlands
O.Y.	Osakeyhtiö	Finland
Plc	Public Limited Company	United Kingdom
Pty	Proprietary Limited Company	Australia
S.A.	Société Anonyme	France
S.A.	Sociedade Anónima	Portugal
S.L.	Sociedad Limitada	Spain
S.p.A	Società per azioni	Italy
S.R.L	Società a Responsabilità Limitata	Italy
sp. z o.o.	Spółka z ograniczoną odpowiedzialnością	Poland
SPRL	Société Privée à Responsabilité Limitée	Belgium
ZRT	Zártkörûen Mûködõ Részvénytársaság	Hungary

inevitably remain many addresses that need coding from an inspection. The coding of laboratories that are part of national or local/regional government usually needs to be done individually, but it is normally clear that they are publicly funded. However there are many private-non-profit organisations whose websites need to be checked in order to determine if they are primarily charities, collecting money in order to carry out or finance research, or voluntary professional associations or research institutes. The former are coded; the latter are not. Research laboratories named after an endowment such as the Wallenberg Foundations in Sweden (there are three) are also given "no code" as the foundation will usually have paid for the building but not for the research unless this is specifically recorded.

The analysis of funding

Once all the funders and addresses listed on a set of papers have been coded (or given "no code"), they are transferred to two worksheets of a special program, whose job is to add the relevant codes to three new columns of a spreadsheet containing details of the papers. These are the reference or index number, the addresses, and the redacted FU column. When the program is run, it generates the explicit codes, the implicit codes and the composite codes. It also lists any addresses or funding organisations that do not appear in the thesauruses so that they can be amended and the program then run again. Codes that appear in both the explicit and implicit columns for a paper are not duplicated in the composite code column. This often occurs with commercial funding when the company is also listed among the paper's addresses.

The funding for each paper is represented by the composite code column, but it needs to be checked. The first step is to use the Excel TRIM function to remove any surplus spaces, and the second step is to use the Excel LEN function to determine the number of characters in the composite code column. It should be either zero or an integer ending in "9". Any composite codes with other numbers of characters will contain a typing error. They can then be checked individually and amended; the corrections should be noted also in the relevant thesaurus. The next step is to count the numbers of funders, F, which is (LEN + 1)/10, and mark in a new column.

The composite codes (now called "combined codes") and F columns are now transferred to a second spreadsheet that also contains the contribution (as fractional counts) to each paper from each of the countries for which the analysis of funding is required in individual columns headed by the country ISO2 digraphs. The second program for funding analysis can now be run. It operates on a double-fractionation basis. This means that the contribution of a funder to a country's research is first fractionated by the presence of the country among the addresses on a paper, and then by the number of funders that could have contributed to that country's research.

Funding organisations are placed in three groups according to their sector and country codes. The first group consists of organisations that only support researchers in their own country. These are governmental and private-non-profit funders with limited exceptions such as the Gates Foundation in the USA, much of whose money is spent overseas. [They are identified by having the country code XN, connoting an international organisation, rather than the country of their location.] There are a few papers where it appears that a charitable funder in one country has supported research elsewhere, but our assumption takes care of the vast majority of cases.

The second group are European organisations, notably the European Commission but also some European professional associations such as the European Society for Medical Oncology (ESMO) and a few charities such as the

Swiss Bridge Foundation. These are given country code EU and are assumed to divide their support between the various European countries listed in the addresses on each paper; their ISO2 codes are listed in the cover sheet of the second program.

The third group are funders who can give support to researchers in any country. They include the WHO and other organs of the United Nations, but also biotech, industrial and pharma companies. In the absence of other information, their support is assumed to go equally to each address.

The program now generates five new spreadsheets labelled as follows: funding alphabetically, funding in order, funding classified, funding calculated and summary. Each of these sheets provides an analysis of the funding of the research of each country whose fractional counts have been tabulated. For example, the "funding in order" lists the largest funders for a country, of which the leading ones are shown in Table 7 for cancer research in Spain. [The names of the organisations in the left column have been added]. In this example, there were 5484 papers with a Spanish address, and Spain's fractional count was 3947, or 72%. Of these, explicit and implicit funding provided support for 2227, or 56.4% of the Spanish output.

Another sheet, "funding calculated" shows the breakdown of all funding by sector and sub-sector, as shown in Table 8. These data can then be used to prepare charts of the sources of support for a country's research. If it is desired to examine the funding for a sub-set of the papers, for example ones on a particular aspect of the disease or type of research, then the procedure with the second funding program should be repeated on a smaller set of papers. This program runs very quickly, so this can readily be done.

The above example is given merely to show the format in which funding output can be tabulated. A full analysis of cancer funding in Europe will be presented and discussed in a subsequent paper.

Table 7: Example of output from second funding program for the leading funders of Spanish cancer research

	ES		
	Papers		5484
	Fractional contribution		3947
	Total funding		2227
(Funding organisation name)	Funders	Group	Papers
Carlos III Health Institute (ISCIII)	ESS-GA-ES	1	498
Spanish Ministry of Education	MEC-GD-ES	1	328
Consejo Superior de Investigaciones Cientificas (CSIC)	CSC-GA-ES	1	154
European Commission	CEC-GD-EU	2	98.3
Catalan Regional Government	CTY-LA-ES	1	94.2
Sanidad y Consumo, Ministero de	HCN-GD-ES	1	88.4
Miscellaneous Spanish PNPs	Y59-NP-ES	1	82.2
Junta de Andalucia (regional government)	JDA-LA-ES	1	57.5
Pfizer Inc.	PFZ-IP-US	3	41.8

Table 8: Example of output from second funding program for the overall analysis of Spanish cancer research funding. For sectoral codes, see Table 1.

ES					
INTL	5484	-BT-	47.6	% GOV	38.5
FRAC	3947	-IN-	49.1	% PNP	8.2
Funded	2227	-IP-	184	% INDY	7.2
-GA-	687	-SN-	1.0	% INTL	2.5
-GD-	447	-SP-	2.9	% NONE	43.6
-LA-	387	INDY	285		
GOV	1521	-GD-EU	99.1		
-CH-	67.9	Other EU	0.0		
-FO-	77.9	-XN	0.0		
-HT-	29.7	INTL	99.1		
-MI-	64.5	NONE	1720		
-NP-	82.2				
PNP	322				

Note: in this table and Table 5 above, numbers > 100 have all been rounded to the nearest integer. The original data can have as many decimal digits as desired.

DISCUSSION

In this paper we have described in some detail the problems attendant on an analysis of research funding using the data contained in the Web of Science, and how we have tackled them. The situation is inevitably complicated by the large number of papers with international collaboration, and the need to make assumptions on how the funding provided by the different sources is distributed among the research partners. Because many papers carry no acknowledgements, and are therefore assumed to have received institutional funding, there is uncertainty over whether this is really distinct from funding from research institutes (departments) that is formally acknowledged. We have assumed that the funding process involves some form of decision-making, and that therefore the proposed research has been specifically recognised as having merit. It follows that research with multiple acknowledgements has been considered several times, and that multiple funding committees have approved it. It should therefore be more meritorious than research with only a single acknowledgement, or none, and should obtain more citations and other measures of impact. Several studies have indeed shown this to be the case.[5,7,8]

The system of coding and analysis that we have described has some limitations. First, the three-letter codes assigned to individual funders may not be enough in the future when many countries develop a multiplicity of funding sources. It may be necessary to change to a four-letter code in order to accommodate up to 456,976 different funders; alternatively we could keep the trigraph codes but allow numbers within the code, e.g., A7J. This would provide for 46,656 codes, which might well be enough for many years. The second limitation is in the attribution of funding by government and private-non-profit funders only to researchers in their country. We know that this is not completely accurate as there are some papers where these types of funders are acknowledged but there are no author addresses within their countries. This is something that needs further investigation, although our assumption is probably good enough for most of them. Third, the two main thesauri are currently held in MS Excel spreadsheets. This is currently satisfactory, but the address thesaurus will at some stage exceed the limit on numbers of rows and a different database will need to be developed.

CONCLUSION

Science research funding is one of the most important global drivers of social, economic and human development. Such funding represents major strategic investments at every level, from supra-nation state (e.g. the European Commission's Horizon 2020 programs), through countries and down to individual charitable funders. Funding in the health sphere is estimated to be about 37 billion USD every year from the top ten funders. However, beyond such aggregate raw analysis little global strategic intelligence exists on science funders and funding.[13] Who funds what research and how, the dynamics of science funding and the gaps as well as saturated domains are all critical aspects of strategic intelligence around science funding needed to inform policy-makers.

The approaches and methods described in this paper provide a way to gain timely, objective and high quality intelligence on science research funders and funding. Such methods will also allow a more detailed and nuanced examination of the socio-technical evolution of science through an economic lens, as well as provide important data for policy.[15] However, it would be of material assistance if research funding organisations could specify the format in which their support should be acknowledged by researchers. Normally, this should be their legal name, but with the country included if unclear, e.g., UK Medical Research Council, Spanish Department of Health. This should be in English if the paper is to be published in that language. Such a requirement could also mandated by journal editors.

ACKNOWLEDGEMENT

This work was supported by a grant from the European Commission under the Seventh Framework programme number EC/FP7/602536.

CONFLICT OF INTEREST

None of the authors have a conflict of interest to declare.

REFERENCES

1. Narin F, Shapiro RT. The extramural role of the NIH as a research support agency. Federation Proceedings. 1977;36(11):2470-6.
2. Dawson G, Lucocq B, Cottrell R, Lewison G. (1998) Mapping the Landscape: National Biomedical Research Outputs 1988-95. London: the Wellcome Trust, Policy Report no 9. ISBN 1869835-95-6.
3. Webster BM. International presence and impact of the UK biomedical research, 1989-2000. Aslib Proceedings. 2005;57(1):22-47.
4. MacLean M, Davies C, Lewison G, Anderson J. Evaluating the research activity and impact of funding agencies. Research Evaluation. 1998;7(1);7-16.
5. Rigby J. Looking for the impact of peer review: does count of funding acknowledgements really predict research impact?. Scientometrics. 2013;94(1):57-73.
6. Wang J, Shapira P. Is there a relationship between research sponsorship and publication impact? An analysis of funding acknowledgements in nanotechnology papers. PLoS ONE. 2015;10(2):e0117227. DOI 10.1371/journal.pone.0117727.
7. Lewison G, Dawson G. The effect of funding on the outputs of biomedical research. Scientometrics. 1998;41(1-2):17-27.
8. Roe PE, Wentworth A, Sullivan R, Lewison G. The anatomy of citations to UK cancer research papers. Proceedings of the 11th Conference on S&T Indicators. Leiden. 2010;225-6.
9. Lewison G, Sullivan R. Conflicts of interest statements on biomedical papers. Scientometrics. 2015;102(3):2151-9.
10. Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: Systematic review. BMJ. 2003;326:1167–70.
11. Paul-Hus A, Desrochers N, Costas R. Characterization, description, and considerations for the use of funding acknowledgement data in Web of Science. Scientometrics, in press: 2016;108(1):167-82. DOI 10.1007/s11192-016-1953-y.
12. Cancer Research UK (2015) Beating Cancer Sooner: Annual Report and Accounts. http://www.cancerresearchuk.org/sites/default/files/annual_report_and_accounts_2014-15.pdf
13. Viergever RF, Hendriks TCC. The 10 largest public and philanthropic funders of health research in the world: what they fund and how they distribute their funds. Health Research Policy and Systems. 2016;14:12. doi:10.1186/s12961-015-0074-z.
14. Røttingen J-A, Regmi S, Eide M, Young AJ, Viergever RF, Årdal C, et al. Mapping available health R&D data: what's there, what's missing and what role for a Global Observatory. Lancet. 2013;382:1286–307. doi: 10.1016/S0140-6736(13)61046-6
15. Sullivan R, Eckhouse S, Lewison G. (2008) Using bibliometrics to inform cancer research policy and spending. In: Monitoring Financial Flows for Health Research 2007: Behind the Global Numbers. Geneva: Global Forum for Health Research: ISBN 978-2-940401-04-8.

How to cite this article: Begum M, Lewison G. Web of Science Research Funding Information: Methodology for its use in Analysis and Evaluation. J Scientometric Res. 2017;6(2):65-73.